BiXFormer: A Robust Framework for Maximizing Modality Effectiveness in Multi-Modal Semantic Segmentation

Jialei Chen*, Xu Zheng, Danda Pani Paudel, Luc Van Gool, Hiroshi Murase, *Life Fellow, IEEE,* Daisuke Deguchi, *Member, IEEE,*

***Abstract*—Multi-modal perception is especially important in scenarios like autonomous driving, where single-modal inputs may fail under challenging conditions. Utilizing multi-modal data enhances scene understanding by providing complementary semantic and geometric information. Existing methods fuse features or distill knowledge from multiple modalities into a unified representation, improving robustness but restricting each modality's ability to fully leverage its strengths in different situations. We reformulate multi-modal semantic segmentation as a mask-level classification task and propose BiXFormer, which integrates Unified Modality Matching (UMM) and Cross Modality Alignment (CMA) to maximize modality effectiveness and handle missing modalities. Specifically, BiXFormer first categorizes multi-modal inputs into RGB and non-RGB modalities, *e.g.*, depth, allowing separate processing for each. This design leverages the well-established pretraining for RGB, while addressing the relative lack of attention to non-RGB modality. Then, we propose UMM, which includes Modality Agnostic Matching (MAM) and Complementary Matching (CM). MAM assigns labels to features from all modalities without considering modality differences, leveraging each modality's strengths. CM then reassigns unmatched labels to remaining unassigned features within their respective modalities, encouraging balanced contributions from available modalities and mitigating the impact of missing modalities. Moreover, to further facilitate UMM, we introduce CMA, which enhances the weaker queries assigned in CM by aligning them with optimally matched queries from MAM. Extensive experiments demonstrate consistent improvements across multiple benchmarks.**

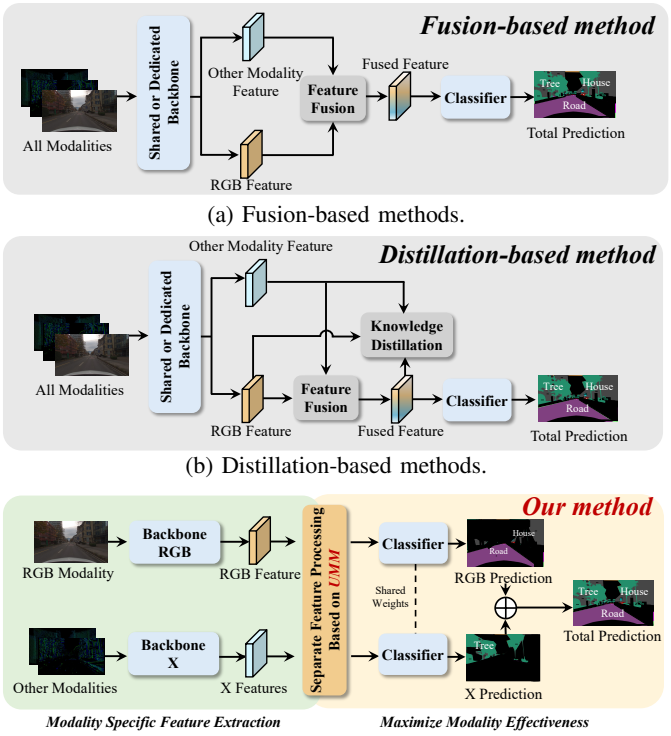
***Index Terms*—Unified Modality Matching, Multi-modality Learning, Semantic Segmentation**

I. Introduction

In multimedia scenarios [1–6], semantic segmentation [7–13], tasked with assigning a class label to each pixel, plays a critical role in enabling scene understanding across diverse environments [1–3]. However, most existing approaches rely solely on RGB images, which often prove inadequate under real-world multimedia conditions such as motion blur, low illumination, or sensor noise, where RGB data may be ambiguous or unreliable [14–16]. To address such limitations, increasing attention has been given to multi-modal semantic segmentation [17–21], which incorporates complementary information from multiple modalities, such as depth, thermal, or event-based data, to enhance segmentation accuracy and robustness across diverse environments.

Existing multi-modal methods often rely on feature fusion [14, 15, 20, 22–28] and knowledge distillation [17–19, 29].

Jialei Chen, Daisuke Deguchi, Hiroshi Murase are with the Graduate School of Informatics, Nagoya University, Nagoya, Japan. Xu Zheng is with AI Thrust, The Hong Kong University of Science and Technology, Guangzhou Campus (HKUST-GZ), Guangzhou, China. Danda Pani Paudel, Luc Van Gool are with the INSAIT, Sofia University, St. Kliment Ohridski. * indicates the corresponding author.



(a) Fusion-based methods.

(b) Distillation-based methods.

(c) Our method.

Fig. 1: Existing methods (fusion-based (a) and knowledge distillation (b)) rely on feature fusion and knowledge distillation, which limits the contribution of individual modalities. In contrast, our method (c) maximizes each modality's (both RGB and non-RGB) contribution by using separate backbones and a unified modality matching mechanism. This allows modalities to focus on the labels they excel at.

Fusion-based methods [14, 15, 20, 22–28], as shown in Fig. 1a, integrate features from multiple modalities through explicitly designed fusion modules to generate robust representations that adapt to diverse scenarios, thereby mitigating the weaknesses of individual modalities. Building upon fusion-based methods, knowledge distillation methods [17–19, 29], illustrated in Fig. 1b, further transfer useful knowledge across modalities to obtain more robust fused representations, *e.g.*, distilling knowledge from stronger modalities to weaker ones [17] or from full-modality settings to partial-modality ones [18]. However, these approaches [14, 15, 17–20, 22–29] face two main limitations, including ① fusion-based [14, 15, 20, 22–28] methods adopt tightly coupled fused features and ***fail to fully exploit the unique strengths of RGB and non-RGB modalities*** for semantic segmentation, leading to suboptimal performance; and ② knowledge distillation methods [17–19, 29] rely heavily

on the teacher model, but selecting an optimal teacher is challenging, as ***fixed teacher models are often not well-suited*** to guide multi-modal learning. Therefore, a new multi-modal framework is needed to fully leverage the strengths of RGB and non-RGB modalities, enabling more robust perception. Additionally, this framework must handle missing modality scenarios, which are common in real-world applications, such as the absence of depth sensors in low-cost systems due to hardware constraints, the unavailability of LiDAR in certain environments like indoor settings, or sensor failures under adverse weather conditions.

To ***maximize RGB and non-RGB modalities' contribution*** and ***mitigate the impact of missing modalities***, we introduce BiXFormer, the ***first*** robust framework designed in a loosely coupled manner, where RGB and non-RGB modalities are processed independently, and their predictions rather than features are combined. This design enables RGB and non-RGB modalities to specialize in the labels they perform best at, leveraging the extensive pretraining available for RGB while compensating for the limited attention given to non-RGB modality. The core idea is shown in Fig. 1c. First, to preserve RGB features when integrating non-RGB modality, we employ two separate backbones: a backbone RGB pre-trained on ImageNet and a backbone X with a modified first layer to handle non-RGB modality. Unlike shared-backbone methods, our approach reduces tightly coupled feature fusion by maintaining distinct backbones for RGB and non-RGB modalities. Additionally, our design enables flexible multi-modal learning with only two backbones, avoiding dedicating backbones for RGB and non-RGB modalities. Second, we introduce Unified Modality Matching (UMM) to ensure effective label matching across modalities. Drawing inspiration from [30], we introduce UMM to assign the most suitable labels to RGB and non-RGB modalities. Concretely, UMM assigns labels through a two-step process: ① **Modality Agnostic Matching** (MAM) first assigns labels to the most relevant queries without modality constraints, ensuring both RGB and non-RGB modalities learn from the categories they are best suited for. However, MAM alone may lead to a dominant modality acquiring most labels, resulting in performance degradation when certain modalities are missing. ② To address this, we introduce **Complementary Matching** (CM), which reassigns unmatched queries to any remaining unassigned labels within RGB and non-RGB modalities, ensuring that missing modality scenarios are effectively handled. Since CM operates on leftover queries, its matching is inherently less optimal than those in MAM. To refine these weaker queries, we introduce Cross Modality Alignment (CMA), aligning weaker queries assigned in CM with the queries matched in MAM.

Overall, BiXFormer avoids tightly coupled feature-level fusion and integrates modality-specific segmentation results at the prediction stage, in contrast to conventional fusion-based models [20, 21] and knowledge distillation–based approaches [17, 19]. Compared to methods that adopt one-to-many matching [31–33], modify the matching metrics [34, 35], or refine queries [36, 37] in a one-step manner, our approach employs a two-step matching process to maximize the contribution of both RGB and non-RGB modalities. BiXFormer achieves consistent improvements across multiple benchmarks. In conclusion, our key contributions are:

(I) We propose **BiXFormer**, a novel robust framework that introduces UMM, maximizing the contributions of both RGB and non-RGB modalities through MAM while mitigating the impact of missing modalities via CM.

(II) We introduce CMA, which leverages the optimally matched queries from MAM to strengthen the weaker queries in CM, improving segmentation robustness.

(III) We evaluate BiXFormer across multiple benchmarks, surpassing existing multi-modal segmentation methods in both full-modality and missing-modality scenarios.

II. Related Works

A. Semantic Segmentation

Semantic segmentation, as one of the most important tasks in computer vision, differs from image classification [38, 39] in that it assigns a class label to each pixel of an image rather than one label to the entire image [9–12, 40, 41]. FCN [7] was the first to treat segmentation as per-pixel classification using CNNs, providing a new perspective for subsequent works. However, FCN suffered from limited receptive field size. With the rapid development of self-attention [42], the receptive field is much enhanced due to the long-range dependency of the transformer. Besides the pixel-level classification view, some methods treat semantic segmentation as a mask-level classification task [30] and achieve impressive performance. Though successful, existing methods only rely on RGB images, which may fall short in some extreme environments, *e.g.*, high-speed movement and low-light conditions. To our knowledge, this is the first attempt to marry the mask-classification segmentation framework with multi-sensor input. Previous multi-modal segmentation models predict pixel-wise labels, whereas BiXFormer produces a set of mask proposals from both RGB and non-RGB modalities and then assigns class labels to them in a coordinated fashion.

B. Multi-modality Learning

In real-world applications, relying on a single modality is often ineffective, particularly under extreme conditions. Consequently, multi-modal learning has gained increasing attention, with numerous research efforts dedicated to this field [16, 17, 20–24, 43–47]. Existing approaches can be broadly categorized into two paradigms: fusion-based methods and knowledge distillation-based methods. Fusion-based methods extract features from RGB and other modalities using either shared [17, 20, 21] or modality-specific [23, 43] backbones. They then integrate these features through various techniques, such as self-attention mechanisms [21], to construct a robust multi-modal representation. On the other hand, knowledge distillation-based methods focus on transferring beneficial knowledge to enhance feature representations [17, 19, 48]. Beyond semantic segmentation, several works [25–28] focus on salient object detection and similarly adopt predominantly fusion-based designs, where non-RGB modality mainly serves as auxiliary information. Specifically, [25, 26] prioritize RGB modality during fusion, where non-RGB inputs mainly provide

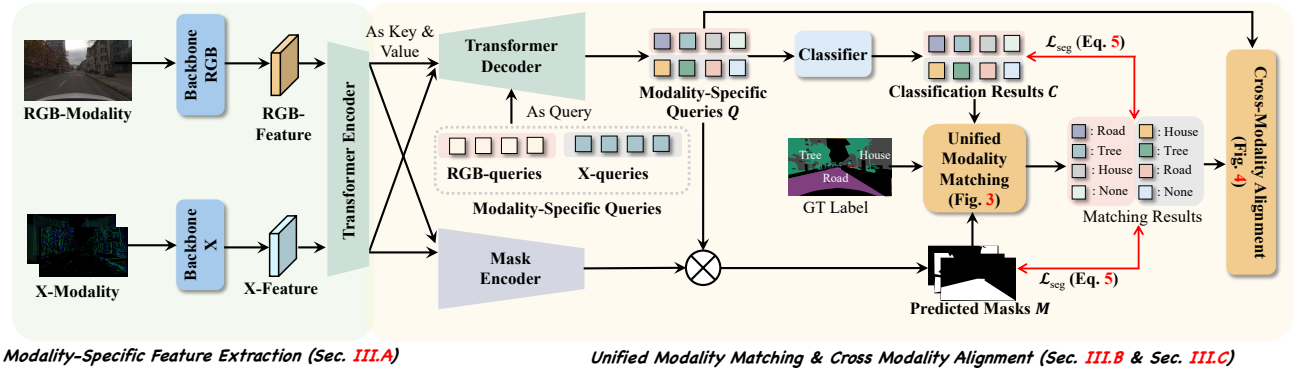


Fig. 2: Overview of our method. Given RGB images and images from X modality, modality-specific features are extracted using their respective backbones: Backbone RGB and Backbone X. These features are then processed by a shared transformer encoder. Modality-specific queries are used as queries, while both RGB and X features serve as keys and values in the Transformer Decoder. The decoded features interact with the Mask Encoder before undergoing Unified Modality Matching (UMM), which assigns labels to modality-specific queries. To further refine the representation, Cross Modality Alignment (CMA) facilitates information transfer between modalities.

complementary cues to enhance RGB modality. [27] further introduces selective fusion across modalities to improve cross-modal interaction, yet the final decoder input still consists of full RGB features combined with selectively aligned non-RGB features. Although [28] treats different modalities more equally during fusion, its fusion modules may suffer from performance degradation when one or more modalities are missing at inference, limiting robustness in practical scenarios.

Despite their effectiveness, these approaches face notable limitations. Fusion-based methods may lead to suboptimal feature representations by failing to fully exploit different modalities. Meanwhile, knowledge distillation-based methods often struggle with selecting an appropriate teacher model, and their representations may not fully capture the richness of each modality. Consequently, we propose BiXFormer, a novel framework that maximizes the contribution of RGB and non-RGB modalities without any pretrained teacher models.

C. Matching Strategies for Query-Based Models

To address the challenge of redundant predictions in object detection, traditional methods rely on Non-Maximum Suppression (NMS) [49] to filter overlapping results. Mask2Former [30] uses a set-based matching mechanism with discrete object queries, enabling the model to directly identify potential objects in an image without the need for post-processing. These queries are assigned labels via Hungarian matching, which optimizes matching based on matching losses. This paradigm has since been extended to segmentation tasks, enabling universal segmentation [30]. However, conventional Hungarian matching often suffers from suboptimal suitability, limiting overall performance. To address this, some methods introduce one-to-many matching [31–33], where a single label is assigned to multiple queries, leveraging the advantages of traditional approaches. Others refine the matching process by modifying matching metrics [34, 35] to ensure stable matching or by enhancing query to improve matching quality [36, 37].

In this paper, we propose Unified Modality Matching (UMM), a novel approach that incorporates two sequential matching steps to maximize the contribution of RGB and non-RGB modalities while preserving the principles of conventional matching. Unlike prior works that primarily focus on refining matching stability or query representation within a single modality, UMM explicitly addresses modality interactions by first performing matching across modalities and then refining matching within RGB and non-RGB modalities.

III. Methods

Preliminary. Our framework processes multi-modal data. Given a dataset $\mathbb{D} = \{\{\mathbf{I}_{ij}\}_{i=1}^{O}, \mathbf{Y}_j\}_{j=1}^{N}$, where $\mathbf{I}_{ij} \in \mathbb{R}^{C_i \times H \times W}$ represents the data from the ith modality of the jth sample, $\mathbf{Y}_j$ is the corresponding pixel-level annotation, H and W denote the height and width of the image, respectively. N indicates the dataset size, C_i indicates the channel number of the ith modality, and O represents the number of modalities. The ground truth set is defined as $\mathbf{Y} = \left\{\mathbf{y}^k = (c_y^k, \mathbf{m}_y^k)\right\}_{k=1}^{U}$ where $c_y^k \in [0, K]$ represents the class label, U indicates the number of unique classes in an image, and $\mathbf{m}_y^k \in [0,1]^{H \times W}$ denotes the corresponding mask. To improve clarity, we define two functions: $C(\mathbf{y})$ extracts the class label c from the annotation $\mathbf{y}$, and $M(\mathbf{y})$ extracts the class-agnostic mask associated with $\mathbf{y}$. For simplicity, we define the first modality as the RGB modality. We follow the data processing pipeline of [20], where the channel of all modalities is set as 3. Besides, the field of view has been aligned for each modality. During training, each batch consists of samples from all modalities.

Method Overview. Our method maximizes the contribution of RGB and non-RGB modalities by processing them separately and assigning suitable labels accordingly. The overall framework is illustrated in Fig. 2. First, we employ two backbones to extract features from RGB and non-RGB modalities. For simplicity, we collectively refer to non-RGB modality as the X modality, as it is treated equally and processed by the same X backbone. Then, to leverage their distinct characteristics, we introduce **Unified Modality Matching (UMM)** for optimal label matching. Finally, to further enhance UMM, we introduce **Cross Modality Alignment (CMA)**, which aligns the best-matching queries from MAM with their suboptimal counterparts

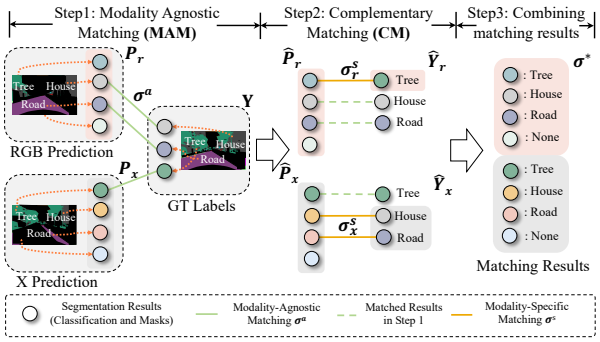


Fig. 3: Overview of the Unified Modality Matching (UMM), where the red dashed line indicates the correspondence between predictions and their respective areas.

in Complementary Matching (CM), enabling information transfer between modalities.

A. Modality Specific Feature Extraction

Existing methods can be broadly categorized into shared-backbone approaches [17, 19] and modality-specific-backbone approaches [45, 46]. Shared-backbone methods use a single ImageNet-pretrained model for all modalities and sometimes incorporate vision-language alignment for RGB. However, they are biased toward RGB features and often underperform on non-RGB inputs due to the lack of dedicated pre-training. Modality-specific methods assign separate backbones to RGB and non-RGB modalities, improving feature extraction but increasing training cost and complexity as the number of modalities grows. To balance these trade-offs, we adopt a dual-backbone design: one backbone for RGB and another shared across all non-RGB (X) modalities. This design leverages the mature pretraining paradigm for RGB (*e.g.*, ImageNet), while compensating for the limited pretraining resources of non-RGB modality, thus supporting more balanced learning.

To maximize the contribution of RGB and non-RGB modalities, we adopt a modality-specific feature extraction strategy. Unlike existing methods [17, 19–21] that employ a single shared backbone across all modalities, we explicitly separate them into RGB and non-RGB (X) modalities. Specifically, the RGB image is denoted as $\mathbf{I}_r = \mathbf{I}_0 \in \mathbb{R}^{3\times H\times W}$, while the X modality is represented as a set of images $\{\mathbf{I}_i\}_{i=1}^{O-1}$. To ensure a comprehensive representation of the X modality, we concatenate them along the channel dimension, forming a unified input $\mathbf{I}_x \in \mathbb{R}^{3(O-1)\times H\times W}$. Instead of using a dedicated backbone for each modality which would lead to a linear increase in parameters with the number of sensors, we use a shared backbone X, leveraging the assumption that generic spatial features can be extracted.

To be specific, we employ two independent backbones: Backbone RGB and Backbone X. The RGB backbone is initialized with an ImageNet-pretrained model and processes $\mathbf{I}_r$ directly. For Backbone X, we also start with an ImageNet-pretrained model but *modify its first convolutional layer to accommodate the concatenated X modality*. Specifically, we replace the original first layer (designed for standard RGB of shape $3 \times C_{\text{out}}$, where C_{out} is the number of output channels) with a randomly initialized layer of shape: $3(O-1) \times C_{\text{out}}$.

Once initialized, Backbone RGB and Backbone X extract dense modality-specific features $\mathbf{F}_r$ and $\mathbf{F}_x$ from $\mathbf{I}_r$ and $\mathbf{I}_x$, respectively. These features are then refined using a **shared transformer encoder**, where self-attention is constrained within RGB and non-RGB modalities to preserve modality-specific representations. Following prior works [30], we formulate semantic segmentation as a ***mask classification task***. The enhanced modality-specific features serve as keys and values in a shared transformer decoder, which processes them using modality-specific queries $\mathbf{Q} = \{\mathbf{Q}_r, \mathbf{Q}_x\}$, where $\mathbf{Q}_r, \mathbf{Q}_x \in \mathbb{R}^{L\times C}$ denote the queries for the RGB and X modalities. Here, L represents the number of queries per modality, and C denotes the feature dimension. To refine $\mathbf{Q}_r$ and $\mathbf{Q}_x$, we apply modality-specific cross-attention:

$$\begin{aligned}\mathbf{Q}_r &= \text{softmax}\left(\frac{f_q(\mathbf{Q}_r) f_k(\mathbf{F}_r)^\top}{\sqrt{C}}\right) f_v(\mathbf{F}_r),\\ \mathbf{Q}_x &= \text{softmax}\left(\frac{f_q(\mathbf{Q}_x) f_k(\mathbf{F}_x)^\top}{\sqrt{C}}\right) f_v(\mathbf{F}_x),\end{aligned} \tag{1}$$

where f_q, f_k, and f_v denote the projection for query, key, and value, respectively. Finally, the modality-specific queries $\mathbf{Q}$ are fed into a classifier to obtain the classification results $\mathbf{C} = \{\mathbf{C}_r, \mathbf{C}_x\}$, where $\mathbf{C}_r, \mathbf{C}_x \in [0,1]^{L\times(K+1)}$ are the classification scores for RGB and X modalities, respectively. Here, $K+1$ accounts for K object classes and an additional "None" class (not matched). Simultaneously, the queries compute the inner product with the features from the mask encoder, generating modality-specific masks $\mathbf{M} = \{\mathbf{M}_r, \mathbf{M}_x\}$, where $\mathbf{M}_r, \mathbf{M}_x \in [0,1]^{L\times H\times W}$ are the predicted segmentation masks from the RGB and X modalities. Finally, we construct the final prediction set $\mathbf{P}$ by combining classification predictions $\mathbf{C}$ and mask predictions $\mathbf{M}$ at the query level. Formally, each prediction is represented as a tuple $(c^i, \mathbf{m}^i)$, where $c^i \in \mathbf{C}$ denotes the predicted class label, and $\mathbf{m}^i \in \mathbf{M}$ is the corresponding mask prediction. The complete set of predictions is $\mathbf{P} = \{(c^i, \mathbf{m}^i) \mid c^i \in \mathbf{C}, \mathbf{m}^i \in \mathbf{M}\}_{i=1}^{2L}$.

To further clarify the design of modality-specific feature extraction, we emphasize that our method does not rely on separate backbone architectures for different modalities. Instead, modality-specific behavior naturally emerges from feature-level complementarity across modalities, together with the proposed UMM and CMA. Since each modality is required to perform semantic segmentation independently, strong high-level semantic discrimination is essential. Following common practice in semantic segmentation (*e.g.*, FCN [7], DeepLab [50], and SETR [51]), we therefore adopt ImageNet pretraining to preserve semantic awareness and ensure stable optimization.

Different modalities mainly vary in their low-level visual characteristics. To accommodate such differences while maintaining high-level semantics, we modify only the first convolutional layer of the X-backbone for input adaptation, while initializing and fine-tuning the remaining layers from ImageNet pretraining. We further conduct complementary experiments in Sec.IV, including: (i) using a single shared backbone across all modalities (Table IV); (ii) removing ImageNet pretraining for the X-backbone (Table X); and (iii) adopting separate backbones for different modalities

(Table VII). Moreover, although both backbones share the same initialization, their representations naturally diverge during training due to modality-dependent inputs, as reflected by the distinct assignment distributions in Fig. 6.

B. Unified Modality Matching (UMM)

To further enhance the contribution of RGB and non-RGB modalities, we propose Unified Modality Matching (UMM), as illustrated in Fig. 3. UMM consists of Modality Agnostic Matching (MAM) and Complementary Matching (CM), ensuring balanced label matching across RGB and X modalities. These steps must be executed sequentially, with MAM first for global matching, followed by CM to refine matching, ensuring optimality across modalities.

We begin with modality agnostic matching, where predictions from both RGB and X modalities are treated as a unified set. The goal of MAM is to assign labels to queries in a globally optimal manner, without considering modality constraints. The optimal matching σ^a is obtained by solving:

$$\sigma^a = \arg\min_{\sigma} \sum_{i=1}^{2L} \mathcal{L}_{match}(\mathbf{P}^{\sigma(i)}, \mathbf{Y}), \tag{2}$$

where the matching loss is defined as:

$$\mathcal{L}_{match}(\mathbf{P}, \mathbf{Y}) = \mathcal{L}_{cls}(\mathbf{C}, \mathbf{Y}) + \mathcal{L}_{mask}(\mathbf{M}, \mathbf{Y}), \tag{3}$$

where $\mathcal{L}_{cls}$ and $\mathcal{L}_{mask}$ indicate the loss function for classification and mask prediction. Class labels ranging from 1 to K represent valid classes, while 0 indicates the "None" category, meaning the prediction is not associated with any ground truth object. For such "None" queries, the corresponding mask is explicitly defined as $\mathbf{0}^{H\times W}$. The obtained matching is $\sigma^a = \{(\mathbf{p}^i, \mathbf{y}_a^i)\}_{i=1}^{2L}$ where $\mathbf{p}^i \in \mathbf{P}$ is the prediction of the ith query, and $\mathbf{y}_a^i$ is the assigned ground truth.

Although this matching step is globally optimal across modalities, it does not account for modality balance, potentially leading to all labels being assigned to a single modality. To this end, we introduce complementary matching. First, we decompose the initial matching σ^a into separate matching for RGB and X modalities $\sigma_r^a = \{(\mathbf{p}_r^i, \mathbf{y}_r^i) \mid \mathbf{p}_r^i \in \mathbf{P}_r\}$, $\sigma_x^a = \{(\mathbf{p}_x^i, \mathbf{y}_x^i) \mid \mathbf{p}_x^i \in \mathbf{P}_x\}$ where $\mathbf{P}_r$ and $\mathbf{P}_x$ indicate the predictions from RGB and X modalities, respectively. By doing so, we extract the set of labels that have already been matched to queries in RGB and X modalities $\mathbf{Y}_r = \{\mathbf{y}_r^i \mid (\mathbf{p}_r^i, \mathbf{y}_r^i) \in \sigma_r^a\}$, $\mathbf{Y}_x = \{\mathbf{y}_x^i \mid (\mathbf{p}_x^i, \mathbf{y}_x^i) \in \sigma_x^a\}$. The remaining unassigned labels for RGB and X modalities are computed as $\hat{\mathbf{Y}}_r = \mathbf{Y} \setminus \mathbf{Y}_r$, $\hat{\mathbf{Y}}_x = \mathbf{Y} \setminus \mathbf{Y}_x$ where $\setminus$ indicates the set minus. Similarly, we define the set of unmatched queries as: $\hat{\mathbf{P}}_r = \{\mathbf{p}_r^i \mid C(\mathbf{y}_r^i) = 0, (\mathbf{p}_r^i, \mathbf{y}_r^i) \in \sigma_r^a\}$, $\hat{\mathbf{P}}_x = \{\mathbf{p}_x^i \mid C(\mathbf{y}_x^i) = 0, (\mathbf{p}_x^i, \mathbf{y}_x^i) \in \sigma_x^a\}$. We perform complementary matching within RGB and X modalities:

$$\begin{aligned} \sigma_r^s &= \arg\min_{\sigma} \sum_{i=1}^{\hat{L}_r} \mathcal{L}_{match}(\hat{\mathbf{P}}_r^{\sigma(i)}, \hat{\mathbf{Y}}_r), \\ \sigma_x^s &= \arg\min_{\sigma} \sum_{i=1}^{\hat{L}_x} \mathcal{L}_{match}(\hat{\mathbf{P}}_x^{\sigma(i)}, \hat{\mathbf{Y}}_x), \end{aligned} \tag{4}$$

where $\hat{L}_r$ and $\hat{L}_x$ represent the number of remaining unassigned predictions in the RGB and X modalities.

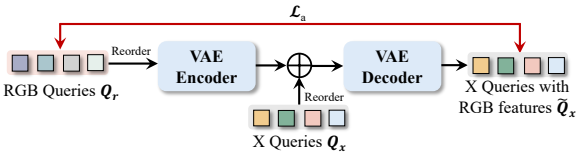


Fig. 4: Overview of cross modality alignment where we only show the process to refine RGB queries for simplicity.

To unify both matching results into a single one, we integrate modality agnostic and modality specific matches. Concretely, to merge the results from modality agnostic and complementary matching into a unified matching σ^*, we first extend σ_r^s and σ_x^s to the same length as σ^a. This is done by ensuring that each query in σ_r^s has a corresponding placeholder in σ_x^s set to "None" and vice versa. Next, for queries already assigned in σ^a, we set their corresponding entries in σ_r^s and σ_x^s to "None". The final matching result is $\sigma^* = \{(\mathbf{p}^i, \mathbf{y}_*^i) \mid \mathbf{p}^i \in \mathbf{P}\}$ where each prediction $\mathbf{p}^i$ is assigned the final label $\mathbf{y}_*^i$. The final label matching is computed as $\mathbf{y}_*^i = (\max(c_a^i, c_{s_r}^i, c_{s_x}^i),\ \max(\mathbf{m}_a^i, \mathbf{m}_{s_r}^i, \mathbf{m}_{s_x}^i))$ where $(c_a^i, \mathbf{m}_a^i)$, $(c_{s_r}^i, \mathbf{m}_{s_r}^i)$, and $(c_{s_x}^i, \mathbf{m}_{s_x}^i)$ denote the class and mask matching from σ^a, σ_r^s, and σ_x^s, respectively. The final segmentation loss is:

$$\mathcal{L}_{seg}(\mathbf{P}, \mathbf{Y}) = \mathcal{L}_{cls}(\mathbf{C}^{\sigma^*}, \mathbf{Y}) + \mathcal{L}_{mask}(\mathbf{M}^{\sigma^*}, \mathbf{Y}), \tag{5}$$

where $\mathbf{C}^{\sigma^*}$ and $\mathbf{M}^{\sigma^*}$ indicate the class and mask prediction under the matching σ^*.

C. Cross Modality Alignment (CMA)

Though effective, the proposed Unified Modality Matching (UMM) framework may result in suboptimal performance due to its query assignment strategy. Specifically, for one category, UMM assigns high-quality queries to one modality while using less effective queries in the other, leading to degraded predictions when combining both modalities. To address this, we propose Cross Modality Alignment (CMA), as illustrated in Fig. 4. CMA leverages the most discriminative queries assigned in MAM to enhance their weaker counterparts in CM, ensuring that both modalities utilize high-quality queries for all categories. By aligning weaker queries with their stronger counterparts across modalities, CMA enables robust features and improves overall segmentation performance.

Given the query sets $\mathbf{Q}_r$ and $\mathbf{Q}_x$, along with the final matching results σ_r^* and σ_x^*, where $\sigma_r^* = \{(\mathbf{p}^i, \mathbf{y}_*^i) \mid \mathbf{p}^i \in \mathbf{P}_r\}$, $\sigma_x^* = \{(\mathbf{p}^i, \mathbf{y}_*^i) \mid \mathbf{p}^i \in \mathbf{P}_x\}$, we first reorder queries based on their assigned class labels, ensuring a consistent indexing structure across modalities. Due to inherent modality differences, directly aligning the queries across modalities may reduce modality discrepancies rather than preserving semantic consistency. To address this, we then introduce a VAE-based refiner to facilitate cross-modal adaptation while explicitly controlling modality discrepancy and semantic alignment. We choose VAE as the refinement model because, unlike GANs, it does not require additional adversarial training steps, making it more stable and efficient. While diffusion models have shown strong generative capabilities, they typically require the target distribution to be relatively static for effective

TABLE I: Comparison with previous methods on real-world benchmark MUSES [16] with three modalities (F: frame camera, E: event camera, L: LiDAR sensor). **Bold** indicates the highest performance while underline indicates the second best performance. Mean refers to the average mIoU across all modality settings.

Method	Backbone	Training Modality	Parameters (M)↓	GFLOPs ↓	Testing Modality (mIoU)							Mean
					F	E	L	FE	FL	EL	FEL	
CMNeXt [20]	Seg-B0 [9]	F	**3.72**	**27.46**	43.37	-	-	-	-	-	-	-
CMX [21]	Seg-B0 [9]	FEL	10.32	62.40	2.52	2.35	3.01	41.15	41.25	2.55	19.30	16.10
CMNeXt [20]			10.30	44.78	3.50	2.77	2.64	6.63	10.28	3.14	46.66	10.80
MAGIC [17]			24.74	444.35	43.22	2.68	22.95	43.51	49.05	22.98	49.02	33.34
Any2Seg [19]			24.73	-	44.40	3.17	22.33	44.51	49.96	22.63	50.00	33.86
Ours	ResNet-18 [38]	FEL	28.91	53.54	62.03	17.20	39.33	57.27	59.56	45.76	62.40	49.07 (15.21↑)
	ResNet-34 [38]		53.75	91.85	66.70	20.60	**46.89**	61.14	68.22	**54.76**	69.46	55.39 (21.53↑)
	ResNet-50 [38]		58.94	100.0	**70.10**	**22.18**	45.57	**64.59**	**69.28**	52.44	**71.36**	**56.50 (22.74↑)**
	Seg-B0 [9]		17.70	59.10	65.76	16.88	39.63	58.42	59.77	49.89	65.81	50.88 (17.02↑)

generation. However, in our method, both the generated target and the input are dynamically changing, making diffusion-based approaches less suitable for our setting. Specifically, the encoder processes RGB queries to extract modality-specific characteristics and infuse them into the X queries, ensuring that essential modality distinctions are retained rather than collapsed. Meanwhile, the decoder refines X query representations by minimizing the discrepancy between modality-specific class centers and their shared semantic counterparts, ensuring that queries from different modalities converge toward the same category representation. The alignment loss is:

$$\mathcal{L}_a = \mathcal{L}_r(\mathbf{Q}_r, \tilde{\mathbf{Q}}_x) + \mathcal{L}_r(\mathbf{Q}_x, \tilde{\mathbf{Q}}_r) + \mathcal{L}_{kld}, \tag{6}$$

where $\mathcal{L}_r$ is a combination of Mean Squared Error (MSE) and Maximum Mean Discrepancy (MMD) loss [52]. $\tilde{\mathbf{Q}}_r$ and $\tilde{\mathbf{Q}}_x$ denote the queries produced by the VAE decoder. $\mathcal{L}_{kld}$ is the standard VAE regularization term, which enforces the variational posterior to match a Gaussian prior.

To further clarify the role of CMA, we provide a variational interpretation based on the Evidence Lower Bound (ELBO), where the expected log-likelihood term preserves semantic correctness under the shared condition $\mathbf{Y}$, while the KL regularization term avoids over-constraining modality-specific variations. Without loss of generality, we assume that $\mathbf{Q}_r$ is the input from a more reliable modality and $\mathbf{Q}_x$ denotes the target-modality queries to be refined. The formulation is symmetric with respect to the choice of input and condition. The objective of CMA is to improve the semantic awareness of $\mathbf{Q}_x$ under the guidance $\mathbf{Q}_r$ with the same labels assigned by UMM. However, rather than directly modeling $p(\mathbf{Q}_x \mid \mathbf{Y})$, which lacks guidance from modality-compatible features, we optimize $p(\mathbf{Q}_x \mid \mathbf{Q}_r, \mathbf{Y})$, where $\mathbf{Q}_r$ provides additional semantic evidence aligned with $\mathbf{Y}$. Directly modeling this conditional distribution is still intractable due to the complex modality gap between $\mathbf{Q}_x$ and $\mathbf{Q}_r$. To address this, we introduce a latent variable $\mathbf{Z}$ to capture modality-specific characteristics that are not shared between $\mathbf{Q}_x$ and $\mathbf{Q}_r$. We then derive ELBO on the conditional log-likelihood:

$$\begin{aligned}\log p(\mathbf{Q}_x \mid \mathbf{Q}_r, \mathbf{Y}) \geq\ & \mathbb{E}_{\mathbf{Z}\sim q_\phi}\left[\log p(\mathbf{Q}_x \mid \mathbf{Z}, \mathbf{Q}_r, \mathbf{Y})\right] \\ & - \mathrm{KL}\left(q_\phi(\mathbf{Z} \mid \mathbf{Q}_r, \mathbf{Y}) \parallel p(\mathbf{Z})\right)\end{aligned} \tag{7}$$

where $p(\mathbf{Z})$ is a standard normal prior. The first term encourages $\mathbf{Q}_x$ to achieve semantic alignment with $\mathbf{Q}_r$ through $\mathcal{L}_r$, while the KL divergence, *i.e.*, $\mathcal{L}_{\mathrm{kld}}$, preserves modality-specific variation within a shared semantic space. This design allows CMA to refine weaker queries using more reliable ones without collapsing modality discrepancies, supporting robust segmentation even under missing-modality scenarios.

D. Training Objectives and Inference

The total loss functions for training are defined as,

$$\mathcal{L} = \mathcal{L}_{seg} + \mathcal{L}_a. \tag{8}$$

During inference, modality-specific queries from the backbone and transformer encoder are passed to the classifier. The classification results are then combined with the predicted masks via the inner product for the prediction. If any sub-modality in X is missing, it is replaced with $\mathbf{0}^{3\times H\times W}$.

IV. Experiments

A. Experimental Setup

Datasets. We evaluate our method on both synthetic and real-world multi-sensor datasets to assess its robustness and generalizability. The MUSES dataset [16] is a real-world driving dataset collected in Switzerland, specifically designed to address challenges arising from adverse visual conditions. It provides multi-sensor recordings from a high-resolution RGB camera, an event camera, and a MEMS LiDAR, along with high-quality 2D panoptic annotations for benchmarking. DELIVER [20] is a synthetic dataset which comprises RGB, depth, LiDAR, and event streams across 25 semantic categories. It encompasses a wide range of environments and includes scenarios involving sensor degradation and failure, making it well-suited for evaluating multi-modal robustness. NYUv2 Dataset. NYUv2 is a standard benchmark for indoor RGB-D semantic segmentation, consisting of 1,449 RGB-D images annotated with 40 semantic categories. The images are captured in diverse indoor scenes using a Kinect sensor. We follow the official split with 795 training and 654 testing images.

Implementation Details. All experiments were conducted on 4 NVIDIA A6000 GPUs. The initial learning rate was set to 6×10^{-5} and adjusted using a polynomial decay strategy with a power of 0.9 over 150 epochs. Additionally, a 10-epoch warm-up phase was applied at 10% of the initial learning rate to stabilize training. AdamW optimizer was employed, and the effective batch size for both datasets was set to 16. Input modality data was cropped to 1024×1024 across benchmarks. The VAE structures for both the encoder and decoder are four layers of Linear-Layernorm-ReLU. We use mean Intersection over Union (mIoU) as the performance metric.

TABLE II: Comparison with previous methods on synthetic benchmark DELIVER [20] with four modalities (R: RGB images, D: depth images, E: Event Camera, L: Lidar sensor) using ResNet-34 [38] as backbone model. **Bold** indicates the highest performance while underline indicates the second best performance. Mean refers to the average mIoU across all modalities.

Method	Testing Modality (mIoU)															Mean
	R	D	E	L	RD	RE	RL	DE	DL	EL	RDE	RDL	REL	DEL	RDEL	
CMNeXt [20]	0.86	0.49	0.66	0.37	47.06	9.97	13.75	2.63	1.73	2.85	59.03	59.18	14.73	59.18	39.07	20.77
MAGIC [17]	32.60	**55.06**	0.52	0.39	**63.32**	33.02	33.12	**55.16**	**55.17**	0.26	**63.37**	**63.36**	33.32	**55.26**	**63.40**	40.49
Ours	**53.32**	50.18	**1.03**	**1.49**	55.58	**53.30**	**53.40**	49.60	51.20	**2.19**	55.57	55.70	**53.77**	54.08	58.29	**43.24** (**2.75**↑)

TABLE III: Comparison with previous methods on NYUv2 [53] with two modalities (RGB and depth) **Bold** indicates the highest performance while underline indicates the second best performance. The number in () indicates the performance without multi-scale inference. Pretrain indicates whether an extra teacher model is used or not.

Model	Backbone	Pretrain	RGB + Depth	RGB	Depth
ACNet [54]	ResNet-50 [38]	-	48.3	-	-
ShapeConv [55]	ResNext-101 [56]	-	51.3	-	-
ESANet [57]	ResNet-34 [38]	-	50.3	-	-
TokenFusion [58]	MiT-B3 [9]	-	54.2	-	-
Omnivore [59]	Swin-B [60]	-	54.0	-	-
CMNeXt [20]	MiT-B4 [9]	-	56.9 (53.6)	38.3 (39.0)	5.7 (5.3)
GOPT [61]	ViT [42]	-	54.3	-	-
Dformer [22]	Dformer-L [22]	✓	**57.2**	-	-
Ours	MiT-B4 [9]	-	54.5 (54.1)	52.6 (52.4)	38.7 (37.7)

TABLE IV: Ablation studies on proposed methods where "O" ,"T", "U", "A" indicate one shared backbone, two separate backbones, UMM, and CMA, respectively.

Method	Testing Modality (mIoU)							Mean
	F	E	L	FE	FL	EL	FEL	
O	24.77	2.99	1.79	24.05	22.65	2.82	24.68	14.82
O + U	34.28	20.30	21.34	35.71	36.64	38.10	43.30	32.66
T + U	**68.35**	**21.23**	45.63	60.89	66.83	50.68	67.70	54.47
T + U + A	66.70	20.60	**46.89**	**61.14**	**68.22**	**54.76**	**69.46**	**55.39**

B. Comparison with Existing Methods

We compare our method with previous approaches on MUSES (real-world) [16] and DELIVER (synthetic) [20] datasets, including missing and full modality compared with fusion-based methods CMNeXt [20] and knowledge-distillation based methods MAGIC [17] during testing. Table I compares our approach with fusion-based and knowledge distillation (KD)-based methods on the MUSES benchmark [16]. Feature fusion methods, such as CMNeXt [20], integrate multi-modal features at the representation level but suffer significant performance drops when trained with only RGB. Additionally, CMX [21] and CMNeXt [20] exhibit severe degradation when individual modalities are used in isolation, demonstrating their sensitivity to missing modalities. In contrast, KD-based methods, such as MAGIC [17], rely on teacher models and achieve strong multi-modal performance but suffer from excessive computational costs. Our approach overcomes these limitations by robust multi-modal learning with reduced computational overhead even when certain modalities are missing.

On DELIVER as shown in Table II, our approach achieves the best results in single-modality settings, particularly excelling in RGB (53.32%) and event (1.03%), while maintaining competitive performance in depth and LiDAR. Notably, MAGIC achieves slightly higher mIoU when all modalities are present.

TABLE V: Ablation studies on UMM.

Method		GFLOPs↓	Testing Modality (mIoU)							Mean
MAM	CM		F	E	L	FE	FL	EL	FEL	
✓		91.83	55.17	6.09	6.62	47.02	55.09	7.74	55.50	33.31
	✓	91.85	66.66	3.32	**47.02**	50.75	66.29	35.62	68.44	48.30
✓	✓	91.85	**66.70**	**20.60**	46.89	**61.14**	**68.22**	**54.76**	**69.46**	**55.39**

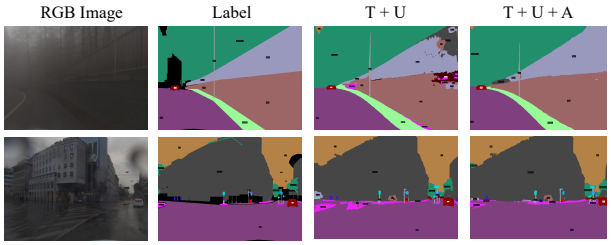


Fig. 5: Visualization between T + U and T + U + A. "T", "U", "A" indicate two separate backbones, UMM, and CMA.

However, in mixed-modality settings, MAGIC's performance degrades significantly if a modality is missing (mean 40.49%), whereas our method maintains strong results.

Table III compares our method with previous approaches on NYUv2 using RGB and depth modalities. Bold and underlined values indicate the best and second-best performances, respectively, and values in parentheses denote results without multi-scale inference. Despite not using a teacher model, our method achieves competitive RGB+Depth performance, comparable to methods with additional supervision. Notably, our model shows strong robustness under missing modality settings, achieving 52.6% (RGB-only) and 38.7% (depth-only).

C. Ablation Studies

To evaluate the effectiveness of the proposed methods, we conduct ablation studies on the MUSES dataset with 200 epochs, using ResNet-34 as the backbone while keeping all hyperparameters unchanged.

Ablations on the proposed modules. Table IV presents the ablation study on key components. A single shared backbone yields poor performance (14.82%), highlighting its limitation in multi-modal settings. Introducing UMM (O+U) brings a clear improvement (32.66%), demonstrating the benefit of unified modality matching. With dual backbones, the performance is further boosted (54.47%), especially under challenging missing-modality settings, indicating the importance of modality-aware feature extraction. Finally, adding CMA (T+U+A) consistently improves the overall performance and achieves the best mean mIoU, with particularly evident gains on non-RGB modality and multi-modality combinations. As visualized in Fig. 5,

TABLE VI: Ablation studies on the refiner in CMA, where "MLP", "M", "C", "Align" denote MLP-based refiner, modality distance, class distance, align both modalities, respectively.

Method	M↑	C↓	Testing Modality (mIoU)							Mean
			F	E	L	FE	FL	EL	FEL	
Align	0.27	2.76	64.54	**21.51**	44.07	59.21	62.66	50.27	63.79	52.29
MLP	0.93	0.57	65.32	20.64	43.47	61.09	62.18	52.13	65.16	52.85
VAE	1.03	0.49	**66.70**	20.60	**46.89**	**61.14**	**68.22**	**54.76**	**69.46**	**55.39**

TABLE VII: Ablation study on modality-specific backbones.

Specific Backbone	Para(M)	GFLOPs	Testing Modality (mIoU)							Mean
			F	E	L	FE	FL	EL	FEL	
✓	72.18	138.0	66.05	**40.86**	40.82	**62.81**	66.05	41.95	52.39	52.99
-	**53.75**	**91.85**	**66.70**	20.60	**46.89**	61.14	**68.22**	**54.76**	**69.46**	**55.39**

CMA produces cleaner and more coherent segmentation maps with fewer spurious predictions, suggesting that CMA better facilitates cross-modal interaction and improves robustness when modality information is incomplete.

Ablations on the unified modality matching. Table V evaluates Unified Modality Matching (UMM) by decomposing it into Modality-Agnostic Matching (MAM) and Complementary Matching (CM). MAM alone performs poorly (33.31%) due to the absence of modality-specific alignment, while CM improves performance (48.30%) but remains suboptimal. Combining both yields the best result (UMM), with only a 0.02 GFLOPs increase, showing that modality-specific matching is both effective and efficient. UMM's two-step matching ensures that no single modality dominates all labels, enhancing robustness under missing modality conditions.

Fig. 6 further analyzes UMM's behavior by showing the category-wise label assignment on the MUSES dataset. Some categories (*e.g.*, fence, traffic sign) favor non-RGB modality, while others (road, pole) prefer RGB. Balanced classes like vegetation demonstrate that UMM adapts matching based on modality strengths, rather than enforcing uniformity.

Ablation studies on the refiner in CMA. In multi-modal segmentation, we define two key metrics: modality distance, measuring distributional divergence between modalities, and class distance, indicating the deviation of RGB and non-RGB features from their shared semantic center. A robust model should maintain distinct modality features (high modality distance) while aligning them semantically (low class distance).

Table VI compares refinement strategies in Cross-Modality Alignment (CMA), using MMD to evaluate modality separation and L1 loss for intra-class consistency. Without a refiner, direct query alignment reduces modality differences (low MMD) but fails to improve semantic consistency (high L1), suggesting that naive alignment collapses modality-specific cues. An MLP refiner improves both metrics modestly, but lacks flexibility for capturing modality variance. Our VAE refiner achieves the best trade-off by learning a latent space that preserves modality distinction and enhances class-wise alignment, leading to superior segmentation performance.

Moreover, CMA does not compromise the alignment established by UMM. In cross-modal semantic segmentation, it is crucial to balance semantic consistency and the preservation of modality-specific characteristics. Directly aligning features across modalities, *e.g.*, $\mathbf{Q}_x$ and $\mathbf{Q}_r$, primarily reduces modality-level discrepancies rather than narrowing the semantic gap. As shown in Table VI, naive feature alignment decreases the modality distance but leads to an increased semantic (class-level) distance, indicating over-smoothing of modality-specific traits and degraded semantic discriminability. In contrast, CMA refines feature distributions under fixed semantic supervision from UMM, while preserving modality-specific diversity through latent space modeling.

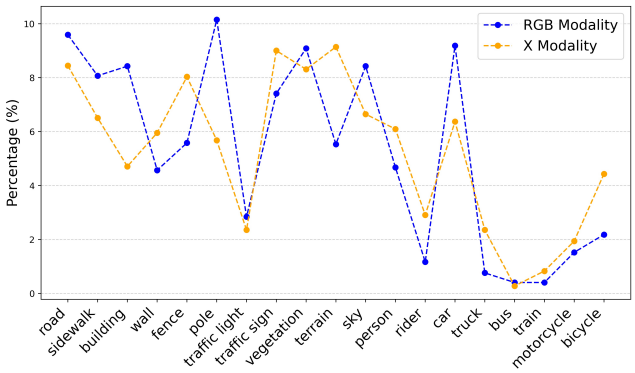


Fig. 6: Label matching distribution under UMM.

TABLE VIII: Ablation studies on backbone structure where "18", "34" indicate the ResNet18 and ResNet34, respectively.

Method		Testing Modality (mIoU)							Mean
RGB	X	F	E	L	FE	FL	EL	FEL	
18	34	65.29	20.1	45.94	63.70	65.80	56.52	68.50	55.12
34	18	66.30	20.4	44.38	64.04	70.07	52.79	68.87	55.26
34	34	**66.70**	20.60	**46.89**	**61.14**	**68.22**	**54.76**	**69.46**	**55.39**

Ablations on the modality specific backbones. Table VII presents an ablation study on modality-specific backbones, where the backbone features of each X modality are averaged for subsequent processing. As shown, introducing modality-specific backbones (3 backbones) increases the model size from 53.75 M to 72.18 M parameters and raises the computational cost from 91.85 to 138.0 GFLOPs. However, this increase in complexity does not translate into better performance, as the mean mIoU decreases from 55.39% to 52.99%. In contrast, the shared X-modality backbone achieves superior performance with significantly lower computational overhead, indicating a more favorable efficiency-accuracy trade-off. These results suggest that a shared-backbone remains effective even under heterogeneous non-RGB modality within our framework.

Ablations on the structures of two backbones. Ablation Study on Backbone Selection. Table VIII presents an ablation study on different backbone configurations, where "18" and "34" denote ResNet-18 and ResNet-34, respectively. The study explores the impact of backbone choices for RGB and X modalities on segmentation performance across different modality settings. The results indicate that using a stronger backbone for the X modality improves performance across most settings. For example, when using ResNet-34 for X while keeping ResNet-18 for RGB, the mean mIoU increases from 55.12% to 55.26%. When both RGB and X use ResNet-34, the model achieves the best performance (55.39% mean mIoU), with noticeable improvements in L (46.89%), FE (61.14%), and FL (68.22%) settings, suggesting that increasing the backbone capacity for the X modality enhances overall segmentation

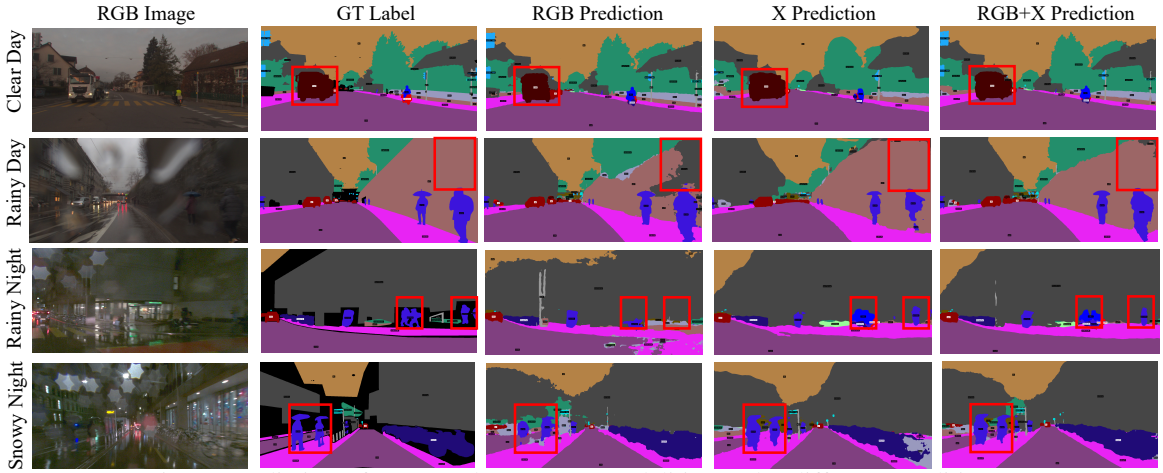


Fig. 7: Prediction of RGB and non-RGB modalities under different conditions.

TABLE IX: Ablation studies on different refinement on CMA.

Method	Testing Modality (mIoU)							Mean
	F	E	L	FE	FL	EL	FEL	
MMD	69.18	20.04	45.63	58.16	66.81	49.07	66.73	53.66
MSE	62.93	18.27	42.17	53.77	61.22	51.62	64.15	50.59
InfoNCE	**70.56**	**20.92**	42.27	**63.48**	65.29	54.79	69.42	55.24
MMD + MSE	66.70	20.60	**46.89**	61.14	**68.22**	**54.76**	**69.46**	**55.39**

TABLE X: Ablation studies on backbone initialization.

Method	Testing Modality (mIoU)							Mean
	F	E	L	FE	FL	EL	FEL	
Random	58.58	**24.98**	9.36	48.75	54.65	31.27	57.11	40.67
ImageNet	**66.70**	20.60	**46.89**	**61.14**	**68.22**	**54.76**	**69.46**	**55.39**

performance, particularly in multi-modal scenarios.

Ablations on the refinement in CMA. Table IX reports an ablation study of loss functions for Cross Modality Alignment (CMA), including MMD, MSE, InfoNCE, and their combinations. MMD achieves moderate performance, suggesting distribution-level alignment helps but lacks instance-level precision. MSE performs worst, indicating direct feature matching is insufficient. InfoNCE yields strong results by enhancing both inter-class separation and intra-class compactness. The best performance is achieved by combining MMD and MSE, showing that integrating global and local alignment objectives provides consistent feature refinement.

Ablation Study on Backbone Initialization. To assess the impact of backbone initialization, we compare models trained from scratch and those initialized with ImageNet weights, as shown in Table X. Pretraining leads to a substantial improvement in overall performance, with mean mIoU increasing from 40.67% to 55.39%. Without pretraining, the model struggles to learn effective representations, particularly for modalities like L and FE. In contrast, ImageNet-initialized models consistently achieve higher scores across all modalities.

D. Qualitative Analysis

Contribution of RGB and X modalities. To evaluate our method's ability to maximize RGB and non-RGB modalities' contribution, we visualize predictions under extreme conditions for RGB (Fig. 7). The results highlight the strengths and limitations of different modalities: RGB and X modalities each specialize in their respective strengths, enabling more accurate segmentation; for example, depth improves boundary delineation, while event cameras better capture dynamic objects.

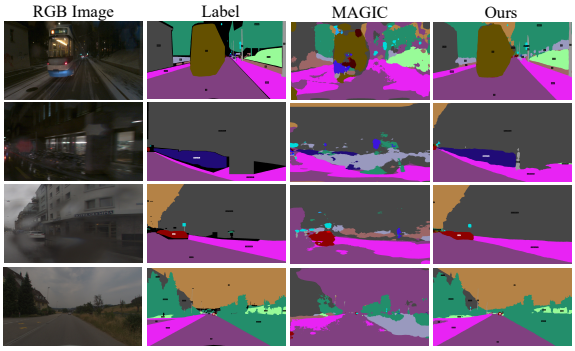


Fig. 8: Predictions visualization compared with MAGIC [17].

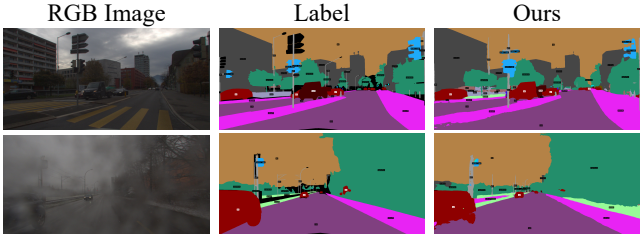


Fig. 9: Predictions visualization of failure cases.

Visualization of our method. Fig. 8 presents a qualitative comparison of semantic segmentation results between our method and MAGIC [17] across diverse driving scenarios, including low-light and adverse weather conditions. Our method demonstrates improved segmentation accuracy, indicating the effectiveness of the proposed method.

Visualization of failure cases. As shown in Fig. 9, although our method achieves strong overall performance, it still exhibits certain semantic and structural inconsistencies. Specifically, large vehicles may be partially misclassified, and individual objects can be fragmented into multiple regions. These observations suggest that further improving semantic coherence represents a promising direction for future work.

V. Conclusion

In this work, we propose **BiXFormer**, a novel robust framework for multi-modal semantic segmentation that effec-

tively leverages the strength of different modalities. Unlike conventional methods that rely on feature fusion or knowledge distillation, BiXFormer preserves modality-specific advantages by reformulating the task as mask-level classification. Our proposed **Unified Modality Matching (UMM)** ensures optimal label matching through a two-step matching process, while **Cross Modality Alignment (CMA)** further refines modality-specific features by aligning strong and weak queries across modalities. By avoiding feature fusion and explicitly addressing missing modality scenarios, BiXFormer achieves significant improvements in segmentation accuracy. Extensive experiments show great performance on several benchmarks.

ACKNOWLEDGMENT

Support for this work was given by the Toyota Motor Corporation (TMC), JSPS KAKENHI Grant Number 23K28164 and JST CREST Grant Number JPMJCR22D1. However, note that this paper solely reflects the opinions and conclusions of its authors and not TMC or any other Toyota entity.

REFERENCES

[1] Q. Ren, Q. Mao, and S. Lu, "Prototypical bidirectional adaptation and learning for cross-domain semantic segmentation," *IEEE Transactions on Multimedia*, vol. 26, pp. 501–513, 2023.

[2] X. Li, Y. Liu, G. Sun, M. Wu, L. Zhang, and C. Zhu, "Towards open-vocabulary video semantic segmentation," *IEEE Transactions on Multimedia*, 2025.

[3] K. Hu, X. Chen, Z. Chen, Y. Zhang, and X. Gao, "Multi-perspective pseudo-label generation and confidence-weighted training for semi-supervised semantic segmentation," *IEEE Transactions on Multimedia*, 2024.

[4] D. Li and S. Du, "Beyond global cues: Unveiling the power of fine details in image matching," in *Proc. IEEE Int. Conf. Multimedia*, 2024, pp. 1–6.

[5] D. Li, Y. Yan, D. Liang, and S. Du, "Msformer: Multi-scale transformer with neighborhood consensus for feature matching," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2023, pp. 1–5.

[6] D. Li and S. Du, "Contextmatcher: Detector-free feature matching with cross-modality context," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 34, no. 9, pp. 7922–7934, 2024.

[7] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, 2015.

[8] L.-C. Chen, Y. Zhu, G. Papandreou, F. Schroff, and H. Adam, "Encoder-decoder with atrous separable convolution for semantic image segmentation," in *Proceedings of the European Conference on Computer Vision*, 2018.

[9] E. Xie, W. Wang, Z. Yu, A. Anandkumar, J. M. Alvarez, and P. Luo, "Segformer: Simple and efficient design for semantic segmentation with transformers," *Advances in Neural Information Processing Systems*, 2021.

[10] J. Chen, D. Deguchi, C. Zhang, X. Zheng, and H. Murase, "Frozen is better than learning: A new design of prototype-based classifier for semantic segmentation," *Pattern Recognition*, 2024.

[11] J. Chen, Z. Quan, C. Zhang, X. Zheng, D. Deguchi, and H. Murase, "Clip-to-seg distillation for zero-shot semantic segmentation," *IEEE Transactions on Circuits and Systems for Video Technology*, 2025.

[12] J. Chen, X. Zheng, D. Li, C. Yi, S. Ito, D. P. Paudel, L. V. Gool, H. Murase, and D. Deguchi, "Split matching for inductive zero-shot semantic segmentation," in *36th British Machine Vision Conference 2025, BMVC 2025, Sheffield, UK, November 24-27, 2025*. BMVA, 2025.

[13] J. Chen, C. Fu, H. Xie, X. Zheng, R. Geng, and C.-W. Sham, "Uncertainty teacher with dense focal loss for semi-supervised medical image segmentation," *Computers in Biology and Medicine*, vol. 149, p. 106034, 2022.

[14] J. Zhang, R. Zhang, L. Xu, X. Lu, Y. Yu, M. Xu, and H. Zhao, "Fastersal: Robust and real-time single-stream architecture for rgb-d salient object detection," *IEEE Transactions on Multimedia*, 2024.

[15] Z. Liu, J. Cheng, J. Fan, S. Lin, Y. Wang, and X. Zhao, "Multi-modal fusion based on depth adaptive mechanism for 3d object detection," *IEEE Transactions on Multimedia*, 2023.

[16] T. Brödermann, D. Bruggemann, C. Sakaridis, K. Ta, O. Liagouris, J. Corkill, and L. Van Gool, "Muses: The multi-sensor semantic perception dataset for driving under uncertainty," in *Proceedings of the European Conference on Computer Vision*, 2024.

[17] X. Zheng, Y. Lyu, J. Zhou, and L. Wang, "Centering the value of every modality: Towards efficient and resilient modality-agnostic semantic segmentation," in *Proceedings of the European Conference on Computer Vision*, 2024.

[18] X. Zheng, H. Xue, J. Chen, Y. Yan, L. Jiang, Y. Lyu, K. Yang, L. Zhang, and X. Hu, "Learning robust anymodal segmentor with unimodal and cross-modal distillation," *arXiv preprint arXiv:2411.17141*, 2024.

[19] X. Zheng, Y. Lyu, and L. Wang, "Learning modality-agnostic representation for semantic segmentation from any modalities," in *Proceedings of the European Conference on Computer Vision*, 2024.

[20] J. Zhang, R. Liu, H. Shi, K. Yang, S. Reiß, K. Peng, H. Fu, K. Wang, and R. Stiefelhagen, "Delivering arbitrary-modal semantic segmentation," in *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, 2023.

[21] H. Liu, J. Zhang, K. Yang, X. Hu, and R. Stiefelhagen, "Cmx: Cross-modal fusion for rgb-x semantic segmentation with transformers," *arXiv e-prints*, pp. arXiv–2203, 2022.

[22] B. Yin, X. Zhang, Z. Li, L. Liu, M.-M. Cheng, and Q. Hou, "Dformer: Rethinking rgbd representation learning for semantic segmentation," *arXiv preprint arXiv:2309.09668*, 2023.

[23] Y. Lv, Z. Liu, and G. Li, "Context-aware interaction network for rgb-t semantic segmentation," *IEEE Transactions on Multimedia*, 2024.

[24] Z. Quan, Q. Chen, M. Zhang, W. Hu, Q. Zhao, J. Hou, Y. Li, and Z. Liu, "Mawkdn: A multimodal fusion wavelet

knowledge distillation approach based on cross-view attention for action recognition," *IEEE Transactions on Circuits and Systems for Video Technology*, 2023.

[25] K. Wang, Z. Tu, C. Li, Z. Liu, and B. Luo, "Unified-modal salient object detection via adaptive prompt learning," *IEEE Transactions on Circuits and Systems for Video Technology*, 2025.

[26] K. Wang, K. Chen, C. Li, Z. Tu, and B. Luo, "Alignment-free rgb-t salient object detection: A large-scale dataset and progressive correlation network," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 39, no. 7, 2025, pp. 7780–7788.

[27] K. Wang, D. Lin, C. Li, Z. Tu, and B. Luo, "Alignment-free rgbt salient object detection: Semantics-guided asymmetric correlation network and a unified benchmark," *IEEE Transactions on Multimedia*, vol. 26, pp. 10 692–10 707, 2024.

[28] K. Wang, Z. Tu, C. Li, C. Zhang, and B. Luo, "Learning adaptive fusion bank for multi-modal salient object detection," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 34, no. 8, pp. 7344–7358, 2024.

[29] Z. Hao, Z. Xiao, Y. Luo, J. Guo, J. Wang, L. Shen, and H. Hu, "Primkd: Primary modality guided multimodal fusion for rgb-d semantic segmentation," in *MM*, 2024.

[30] B. Cheng, I. Misra, A. G. Schwing, A. Kirillov, and R. Girdhar, "Masked-attention mask transformer for universal image segmentation," in *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, 2022.

[31] Q. Chen, X. Chen, J. Wang, S. Zhang, K. Yao, H. Feng, J. Han, E. Ding, G. Zeng, and J. Wang, "Group detr: Fast detr training with group-wise one-to-many assignment," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023.

[32] Z. Zong, G. Song, and Y. Liu, "Detrs with collaborative hybrid assignments training," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023.

[33] D. Jia, Y. Yuan, H. He, X. Wu, H. Yu, W. Lin, L. Sun, C. Zhang, and H. Hu, "Detrs with hybrid matching," in *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, 2023.

[34] S. Liu, T. Ren, J. Chen, Z. Zeng, H. Zhang, F. Li, H. Li, J. Huang, H. Su, J. Zhu *et al.*, "Detection transformer with stable matching," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023.

[35] Y. Pu, W. Liang, Y. Hao, Y. Yuan, Y. Yang, C. Zhang, H. Hu, and G. Huang, "Rank-detr for high quality object detection," *Advances in Neural Information Processing Systems*, 2024.

[36] Z. Hu, K. Gao, X. Zhang, J. Wang, H. Wang, Z. Yang, C. Li, and W. Li, "Emo2-detr: Efficient-matching oriented object detection with transformers," *IEEE TGRS*, 2023.

[37] J. Zhao, F. Wei, and C. Xu, "Hybrid proposal refiner: Revisiting detr series from the faster r-cnn perspective," in *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, 2024.

[38] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, 2016.

[39] C. Zhu, B. Xiao, L. Shi, S. Xu, and X. Zheng, "Customize segment anything model for multi-modal semantic segmentation with mixture of lora experts," *arXiv preprint arXiv:2412.04220*, 2024.

[40] X. Zheng, J. Zhu, Y. Liu, Z. Cao, C. Fu, and L. Wang, "Both style and distortion matter: Dual-path unsupervised domain adaptation for panoramic semantic segmentation," in *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, 2023, pp. 1285–1295.

[41] X. Zheng, P. Zhou, A. V. Vasilakos, and L. Wang, "Semantics distortion and style matter: Towards source-free uda for panoramic segmentation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 27 885–27 895.

[42] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly *et al.*, "An image is worth 16x16 words: Transformers for image recognition at scale," *arXiv preprint arXiv:2010.11929*, 2020.

[43] H. Cao, Z. Zhang, Y. Xia, X. Li, J. Xia, G. Chen, and A. Knoll, "Embracing events and frames with hierarchical feature refinement network for object detection," in *Proceedings of the European Conference on Computer Vision*, 2024.

[44] H. Maheshwari, Y.-C. Liu, and Z. Kira, "Missing modality robustness in semi-supervised multi-modal semantic segmentation," in *WACV*, 2024.

[45] S. Woo, S. Lee, Y. Park, M. A. Nugroho, and C. Kim, "Towards good practices for missing modality robust action recognition," in *AAAI*, 2023.

[46] Y. Zhang, C. Peng, Q. Wang, D. Song, K. Li, and S. K. Zhou, "Unified multi-modal image synthesis for missing modality imputation," *TMI*, 2024.

[47] H. Liu, D. Wei, D. Lu, J. Sun, L. Wang, and Y. Zheng, "M3ae: multimodal representation learning for brain tumor segmentation with missing modalities," in *AAAI*, 2023.

[48] X. Zheng, Y. Lyu, L. Jiang, J. Zhou, L. Wang, and X. Hu, "Magic++: Efficient and resilient modality-agnostic semantic segmentation via hierarchical modality selection," *arXiv preprint arXiv:2412.16876*, 2024.

[49] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár, "Focal loss for dense object detection," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2017.

[50] L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille, "Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2017.

[51] S. Zheng, J. Lu, H. Zhao, X. Zhu, Z. Luo, Y. Wang, Y. Fu, J. Feng, T. Xiang, P. H. Torr *et al.*, "Rethinking semantic segmentation from a sequence-to-sequence perspective with transformers," in *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, 2021.

[52] J. Chen, D. Deguchi, C. Zhang, and H. Murase, "Gen-

eralizable semantic vision query generation for zero-shot panoptic and semantic segmentation," *arXiv preprint arXiv:2402.13697*, 2024.

[53] N. Silberman, D. Hoiem, P. Kohli, and R. Fergus, "Indoor segmentation and support inference from rgbd images," in *Proceedings of the European Conference on Computer Vision*. Springer, 2012, pp. 746–760.

[54] X. Hu, K. Yang, L. Fei, and K. Wang, "Acnet: Attention based network to exploit complementary features for rgbd semantic segmentation," in *2019 IEEE international conference on image processing (ICIP)*. IEEE, 2019, pp. 1440–1444.

[55] J. Cao, H. Leng, D. Lischinski, D. Cohen-Or, C. Tu, and Y. Li, "Shapeconv: Shape-aware convolutional layer for indoor rgb-d semantic segmentation," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2021, pp. 7088–7097.

[56] S. Xie, R. Girshick, P. Dollár, Z. Tu, and K. He, "Aggregated residual transformations for deep neural networks," in *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, 2017, pp. 1492–1500.

[57] D. Seichter, M. Köhler, B. Lewandowski, T. Wengefeld, and H.-M. Gross, "Efficient rgb-d semantic segmentation for indoor scene analysis," in *2021 IEEE international conference on robotics and automation (ICRA)*. IEEE, 2021, pp. 13 525–13 531.

[58] Y. Wang, X. Chen, L. Cao, W. Huang, F. Sun, and Y. Wang, "Multimodal token fusion for vision transformers," in *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, 2022, pp. 12 186–12 195.

[59] R. Girdhar, M. Singh, N. Ravi, L. Van Der Maaten, A. Joulin, and I. Misra, "Omnivore: A single model for many visual modalities," in *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, 2022, pp. 16 102–16 112.

[60] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, and B. Guo, "Swin transformer: Hierarchical vision transformer using shifted windows," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021.

[61] Q. He, "Prompting multi-modal image segmentation with semantic grouping," in *Proceedings of the AAAI conference on artificial intelligence*, vol. 38, no. 3, 2024, pp. 2094–2102.

JIALEI CHEN (Student Member, IEEE) received the B.Eng. and M.Eng. degrees from Northeastern University, Shenyang, China in 2019 and 2022. He is currently pursuing a Ph.D. degree in information science from Nagoya University, Japan. His main research interests include semantic segmentation, zero-shot learning, and image processing.

Xu Zheng (Student Member, IEEE) is a Ph.D. student in the Visual Learning and Intelligent Systems Lab, Artificial Intelligence Thrust, The Hong Kong University of Science and Technology, Guangzhou Campus (HKUST-GZ). He got his B.E. and M.S. from Northeastern University, China. His research interests include multi-modal learning, sensing and perception techniques.

Danda Pani Paudel (Member, IEEE) received the master's and Ph.D. degrees in computer vision from the University of Burgundy, Dijon, France, in 2012 and 2015, respectively. He is a faculty member at INSAIT, Sofia University. He was a Post-Doctoral Researcher with the Computer Vision Laboratory, ETH Zürich, Zürich, Switzerland. He is currently working in the field of image and video enhancement using weak supervision methods. His current research interests include low-level vision, unsupervised learning, and dynamic scene reconstruction and understanding.

Luc Van Gool is a full professor for Computer Vision at ETH Z¨ urich, the KU Leuven and INSAIT. He leads research and/or teaches at all three institutions. He has authored over 900 papers. He has been a program committee member of several major computer vision conferences (e.g. Program Chair ICCV'05, Beijing, General Chair of ICCV'11, Barcelona, and of ECCV'14, Z¨ urich). His main interests include 3D reconstruction and modeling, object recognition, and autonomous driving. He received several best paper awards (e.g. David Marr Prize '98, Best Paper CVPR'07). He received the Koen-derink Award in 2016 and the "Distinguished Researcher" nomination by the IEEE Computer Society in 2017. In 2015 he also received the 5-yearly Excellence Prize by the Flemish Fund for Scientific Research. He was the holder of an ERC Advanced Grant (VarCity). Currently, he leads computer vision research for autonomous driving in the context of the Toyota TRACE labs at ETH and in Leuven, and has an extensive collaboration with Huawei on the topic of image and video enhancement.

HIROSHI MURASE (Life Fellow, IEEE) received the B.Eng., M.Eng., and Ph.D. degrees in electrical engineering from Nagoya University, Japan. In 1980, he joined Nippon Telegraph and Telephone Corporation (NTT). From 1992 to 1993, he was a Visiting Research Scientist with Columbia University, New York. Since 2003, he has been a Professor with Nagoya University. Since 2021, he has been an Emeritus Professor. His research interests include computer vision, pattern recognition, and multimedia information processing. He is a fellow of the IPSJ and the IEICE. He was awarded the IEEE CVPR Best Paper Award, in 1994, the IEEE ICRA Best Video Award, in 1996, the IEICE Achievement Award, in 2002, the IEEE Multimedia Paper Award, in 2004, and the IEICE Distinguished Achievement and Contributions Award, in 2018. He received the Medal with Purple Ribbon from the Government of Japan, in 2012.

DAISUKE DEGUCHI (Member, IEEE) received the B.Eng. and M.Eng. degrees in engineering and the Ph.D. degree in information science from Nagoya University, Japan, in 2001, 2003, and 2006, respectively. He became a Postdoctoral Fellow at Nagoya University, in 2006. From 2008 to 2012, he was an Assistant Professor at the Graduate School of Information Science. From 2012 to 2019, he was an Associate Professor at the Information Strategy Office. Since 2020, he has been an Associate Professor with the Graduate School of Informatics. He is working on object detection, segmentation, recognition from videos, and their applications to ITS technologies, such as detection and recognition of traffic signs. He is a member of IEICE and IPS Japan.